联合特征选择和光滑表示的子空间聚类算法

郑建炜 路 程 秦梦洁 陈婉君

摘 要 基于自表示关联图的谱聚类模型性能受冗余特征影响较大.为了缓解高维数据无效特征的负面影响,文中提出联合特征选择和光滑表示的子空间聚类算法.首先基于自表示思想构建系数矩阵,将特征选择与数据重构纳入同一框架,同时使用权值因子衡量相关特征贡献度,并对系数矩阵进行组效应约束以保持局部性.通过交替变量更新法优化目标函数模型.在人造数据与标准数据库上的实验表明,文中算法在各项性能上均较优.

关键词 子空间聚类,自表示,特征选择,光滑表示,组效应

引用格式 郑建炜,路 程,秦梦洁,陈婉君.联合特征选择和光滑表示的子空间聚类算法.模式识别与人工智能,2018,31(5):409-418.

DOI 10.16451/j.cnki.issn1003-6059.201805003 **中图法分类号** TP 391

Subspace Clustering via Joint Feature Selection and Smooth Representation

ZHENG Jianwei, LU Cheng, QIN Mengjie, CHEN Wanjun

ABSTRACT The performance of self-representation based methods is affected by redundant high-dimensional features. Therefore, a subspace clustering method via joint feature selection and smooth representation (FSSR) is proposed in this paper. Firstly, the idea of feature selection is integrated into the self-representation based coefficient matrix learning framework. Meanwhile, a weight factor is adopted to measure different contributions of correlated features. Furthermore, a group effectiveness constraint is imposed on the coefficient matrix for the preservation of locality property. An alternating direction method of multipliers (ADMM) based algorithm is derived to optimize the proposed cost function. Experiments are conducted on synthetic data and standard databases and the results demonstrate that FSSR outperforms the state-of-the-art approaches in both accuracy and efficiency.

Key Words Subspace Clustering, Self-representation, Feature Selection, Smooth Representation, Grouping Effect

Citation ZHENG J W, LU C, QIN M J, CHEN W J. Subspace Clustering via Joint Feature Selection and Smooth Representation. Pattern Recognition and Artificial Intelligence, 2018, 31(5): 409-418.

收稿日期: 2017-10-09; **录用日期:** 2018-01-08
Manuscript received October 9, 2017;
accepted January 8, 2018
国家自然科学基金项目(No.61602413)、**浙江省自然科学基金项目**(No.LY15F030014) **资助**
Supported by National Natural Science Foundation of China (No. 61602413), Natural Science Foundation of Zhejiang Province (No.LY15F030014)
本文责任编委 王士同
Recommended by Associate Editor WANG Shitong
浙江工业大学 计算机科学与技术学院 杭州 310014
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310014

在计算机视觉、图像处理、多媒体和数据挖掘等现实应用中,观测到的数据集通常源自于多个子空间的联合分布.因此,随之产生的子空间聚类(Subspace Clustering, SC)问题引起广泛关注.子空间聚类是为每个数据点寻找一个低维子空间进行拟合,将每个数据样本分别集中到各自的子空间中.在众多 SC 求解方案中,谱聚类[1-2]算法以谱分解操作[3]为基本思路,将 SC 问题转化为无向连接图的 K 均值聚类问题,具有扎实的理论基础和广泛的应用前景.

谱聚类成功的关键在于关联矩阵(Affinity Matrix)[4]的构造,常见方法有相似图法和表示系数

法.相似图法包括 ε 邻域图(ε-Neighborhood Graph)、k 近邻图(k-Nearest Neighbor Graph)、全连接图(Fully Connected Graph)等,该方法在参数敏感性、鲁棒性和数据适用性等方面存在明显缺陷.基于表示系数构建关联矩阵的方法有:稀疏表示(Sparse Representation, SR)[5]、低秩表示(Low-Rank Representation, LRR)[6]和光滑表示(Smooth Representation, SMR)[7].

Yin 等[8]提出拉普拉斯正则化低秩表示聚类(Laplacian Regularized LRR, LRLRR),使用流形结构的拉普拉斯项,改善低秩表示局部性较差的弱点.Nie 等[9]提出分组结构低秩模型(Group Structure constraint Low-Rank model, GSLR),使用 Schatten p 范数代替 LRR 中的迹范数,低秩近似效果良好,但存在鲁棒性较弱和运行效率较低等缺点.Hu 等[7]提出光滑表示聚类(Smooth Representation Clustering, SRC),是一种性能较优的聚类算法,它依据相似样本的表示系数之间具有的组效应进行聚类.Guo[10]提出联合学习数据表示及其关联矩阵(Data Representations and Affinity Matrix, DRAM)的子空间聚类,使用表示系数构造关联矩阵,取得较优性能,但关联矩阵的实际含义严重偏离其初始概念,过多的参数限制其扩展性.上述方法都是在原始的特征空间中建立数据线性组合关系,同时,使用非相关特征、噪声或冗余特征,因此增加算法复杂度,降低学习性能.

从原始特征空间中选择能保持数据本质结构的特征子空间,称为特征选择(Feature Selection, FS).无监督特征选择通常根据样本相似度、流形结构和鉴别性信息等指标获得相应的特征子空间.近些年,无监督特征选择[11]开始用于针对高维特征数据的子空间聚类中,它通过对特征学习矩阵进行稀疏约束,选择非零行对应的特征子空间.Zhu 等[12]提出无监督特征选择引导的子空间聚类算法(Subspace Clustering Guided Unsupervised FS, SCUFS),学习具有数据子空间结构的全局相似图,迭代生成关联矩阵和伪标签,提高伪标签的辨识度.Wang 等[13]提出凹型约束特征选择(Concave Regularization FS, CRFS),使用泛化的 $l_{2,p}$ 范数约束特征选择矩阵,当 $0 \leqslant p < 1$ 且 p 值较小时,$l_{2,p}$ 范数的稀疏性强于 $l_{2,1}$ 范数.Yan 等[14]提出局部保持分数特征选择(FS via Locality Preserving Score, FSLPS),以局部鉴别信息作为评价准则,选择样本相似度保持得分最高的特征子集.上述基于特征选择的子空间聚类方法通常先选出特征子集,然后使用所选特征进行数据分簇[15],这样不仅隔离特征鉴别信息与聚类结果的相互反馈,也未有效降低算法复杂度.

针对上述问题,本文提出联合特征选择和光滑表示(Joint Feature Selection and Smooth Representation, FSSR)的子空间聚类算法,将特征选择融入基于自表示的谱聚类算法,在构造系数矩阵的同时考虑样本之间的特征相关性.为了使特征选择更准确,使用特征权值这一概念对相关特征进行贡献度分析.此外,使用具有组效应的迹范数约束系数矩阵,增强块对角化分布状态.在人造数据集和标准数据集上的实验表明,本文算法的各项性能较优.

1 相关工作

1.1 自表示

基于数据表示的子空间聚类通过计算样本表示系数构造关联矩阵,性能良好.表示系数矩阵揭示数据潜在的子空间属性,可在一定条件下形成块对角分布[16];反之,如果表示系数矩阵呈现块对角分布,可据此挖掘数据的子空间属性.

给定输入样本集

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \mathbf{R}^{d \times n},$$

具有 n 个样本,d 维特征,该样本集可划分为 k 个相互正交的子集,即 $\cup_{i=1}^{k} \boldsymbol{S}_i$,$\boldsymbol{S}_i$ 表示低维子空间.为使每个样本 $\boldsymbol{x}_i$ 聚集到相应的子空间 $\boldsymbol{S}_\alpha$($\alpha = 1, 2, \cdots, k$),定义一个系数矩阵 $\boldsymbol{Z} \in \mathbf{R}^{n \times n}$,对于样本 $\boldsymbol{x}_i \in \boldsymbol{S}_\alpha$ 可表示为其它样本的线性组合:

$$\boldsymbol{x}_i = \sum_{j \neq i} Z_{ij} \boldsymbol{x}_j.$$

当 $\boldsymbol{x}_i \notin \boldsymbol{S}_\alpha$ 时,对应的 $Z_{ij} = 0$.对于整个样本集,有 $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}$.

受噪声或奇异样本的干扰,数据自表示可进一步描述为 $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}$,其中 $\boldsymbol{E}$ 为噪声或奇异样本.通常,求解最优系数矩阵 $\boldsymbol{Z}^*$ 如下:

$$\begin{aligned} &\min_{\boldsymbol{Z}} \alpha \| \boldsymbol{X} - A(\boldsymbol{X})\boldsymbol{Z} \|_l + \lambda \Omega(\boldsymbol{X}, \boldsymbol{Z}), \\ &\text{s.t. } \boldsymbol{Z} \in C, \end{aligned} \tag{1}$$

其中:$A(\boldsymbol{X})$ 为一种字典,通常根据先验知识确定;$\boldsymbol{X} - A(\boldsymbol{X})\boldsymbol{Z}$ 称为误差项或保真项,衡量重构数据 $A(\boldsymbol{X})\boldsymbol{Z}$ 与数据 $\boldsymbol{X}$ 之间的近似程度.针对不同的噪声分布,选择不同的范数模型约束误差;$\Omega(\boldsymbol{X}, \boldsymbol{Z})$ 称为惩罚项或正则项,对系数矩阵 $\boldsymbol{Z}$ 和数据 $\boldsymbol{X}$ 进行不同类型的约束;C 为系数矩阵 $\boldsymbol{Z}$ 的约束条件;α 为误差项的参数;λ 为惩罚项的参数.

各种自表示算法的区别在于保真项范数模型 $\|\cdot\|_l$、惩罚项 $\Omega(X,Z)$ 和约束条件 C 形式的不同,表 1 给出本文出现的自表示算法中各子项的实施形式,其中 λ 和 β 为惩罚项参数,$\varnothing$ 为空集,S 为相似度矩阵,由系数矩阵 Z 计算得到. $\|\cdot\|_1$ 表示 l_1 范数,通常用于含有拉普拉斯(稀疏)噪声的数据[17]; $\|\cdot\|_{2,1}$ 表示 $l_{2,1}$ 范数,可以有效抑制样本特征的离群点[18]; l_2 范数对高斯噪声数据具有较好的抑制作用[19]; $\|\cdot\|_F$ 表示 Frobenious 范数,适用于高斯噪声分布的数据; $\|\cdot\|_*$ 表示核范数,通常作为低秩函数的凸近似.

表 1　常见的自表示算法实施形式

Table 1　Implementation of self-representation algorithms

算法	$\|\cdot\|_l$	$\Omega(X,Z)$	C
SSC[5]	$\|\cdot\|_1$	$\|Z\|_1$	$\{Z \mid z_{ii}=0\}$
LRR[6]	$\|\cdot\|_{2,1}$	$\|Z\|_*$	$\varnothing$
LRLRR[8]	$\|\cdot\|_1$	$\|Z\|_* + \lambda\|Z\|_1 + \beta\,\mathrm{tr}(ZLZ^{\mathrm{T}})$	$\{Z \mid z_{ij}\geqslant 0\}$
SRC[7]	$\|\cdot\|_F^2$	$\mathrm{tr}(ZLZ^{\mathrm{T}})$	$\varnothing$
DRAM[10]	$\|\cdot\|_F^2$	$\|Z\|_F^2 + \lambda\|S\|_F^2 + \beta\,\mathrm{tr}(ZL_\alpha Z^{\mathrm{T}})$	$\{S \mid s_{ij}\geqslant 0\}$
LSR[20]	$\|\cdot\|_F^2$	$\|Z\|_F^2$	$\varnothing$
CASS[21]	$\|\cdot\|_F^2$	$\sum_i \|X\mathrm{diag}(z_i)\|_*$	$\varnothing$

1.2　组效应

Xu 等[22]指出,若 2 个样本在数据空间上相近,表示系数也相近,进而提出组效应(Grouping Effect)的概念.

定义 1 组效应　给定数据 X,对于 $i\neq j$,z_i、z_j 分别为 x_i、x_j 的表示系数,$\|x_i-x_j\|_2\to 0$ 时,则有 $\|z_i-z_j\|_2\to 0$ 成立,则其系数矩阵 Z 具有组效应.

根据定义 1,Lu 等[20]提出的最小二乘回归(Least Squares Regression, LSR)和相关性自适应子空间分割(Correlation Adaptive Subspace Segmentation, CASS)[21]模型都具有组效应. 强制组效应(Enforced Grouping Effect, EGE)理论[7]给出一组可使系数矩阵具有组效应的条件.

定义 2 EGE **条件**　关于式(1)的 EGE 条件有

1) $A(X)$ 关于 X 连续,$\Omega(X,Z)$ 关于 X 连续,$Z\in C$;

2) 式(1)有唯一解 Z^*,且 Z^* 不是 C 中的孤立点;

3) 对于任意置换矩阵 P,$Z\in C$ 当且仅当 $ZP\in C$,且 $\Omega(X,Z)=\Omega(XP,Z)$;

4) 对于任意置换矩阵 P,$A(XP)=A(X)P$,$Z\in C$ 当且仅当

$$P^{\mathrm{T}}ZP\in C,\ \Omega(X,Z)=\Omega(XP,P^{\mathrm{T}}ZP).$$

可以证明,如果式(1)满足 EGE 条件 1) ~ 3),其最优解 Z^* 具有组效应. 如 LRR[6]的目标问题满足 EGE 条件 1)、3),且 LRR 可证明具有唯一的最优解,则 LRR 具有组效应. 根据 EGE 条件,式(1)中有如下形式的 $\Omega(Z)$ 和 C:

$$\Omega(Z)=\sum_{j=1}^{n}\left(\sum_{i=1}^{n}|Z_{ij}|^p\right)^q,\ p>1,\ q\geqslant\frac{1}{p},\ C=\varnothing;$$

$$\Omega(Z)=\mathrm{tr}((ZHZ^{\mathrm{T}})^p),\ H\succ 0,\ p\geqslant 1/2,\ C=\varnothing;$$

$$\Omega(Z)=\mathrm{tr}((Z^{\mathrm{T}}HZ)^p),\ H\succ 0,\ p\geqslant 1/2,\ C=\varnothing.$$

则其表示系数矩阵具有组效应.

2　联合特征选择和光滑表示的子空间聚类算法

2.1　目标问题设计

针对现存基于特征选择的子空间聚类方法两步操作的缺点,本文使用特征选择后的数据进行自表示,将无监督特征选择和数据分簇整合为一步操作. 为了运算简单,使用给定数据集 X 作为学习字典. 特征学习选择 $M(M\leqslant d)$ 个特征进行信息标记,定义一个特征选择向量

$$p=[p_1,p_2,\cdots,p_d]\in\mathbf{R}^d,\ \|p\|_0=M.$$

若第 i 个特征为相关特征,$p_i=1$,表示选择该特征;否则,$p_i=0$,表示未选择该特征. 经过特征选择的数据

$$x_i^p=[p_1x_{i1},p_2x_{i2},\cdots,p_dx_{id}]=\mathrm{diag}(p)x_i,$$

其中,$\mathrm{diag}(p)\in\mathbf{R}^{d\times d}$ 是以向量 p 为对角线的对角矩阵. 数据 X 可以表示为

$$\mathrm{diag}(p)X=\mathrm{diag}(p)XZ+E,$$

由此可得误差项

$$\min_{Z,p}\frac{1}{2}\|\mathrm{diag}(p)(X-XZ)\|_F^2$$
$$\text{s.t.}\quad p\in\{0,1\},\ \|p\|_0=M.$$

上式虽然选择相关特征,但仅对特征是否相关进行简单判定,不能准确区分不同特征的重要程度. 鉴于此种情况,进一步引入特征权值向量 $p\in[0,1]^d$ 代替特征选择向量,p 的取值不再限于 0 和 1,而是赋予各个特征不同的权值,具有明确的相关性鉴别意义. 因此,融合特征权值的自表示模型的误差项形式如下:

$$\min_{\boldsymbol{Z},\boldsymbol{p}} \frac{1}{2}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XZ})\|_{\mathrm{F}}^2,$$

$$\text{s.t.}\quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1,\ p_i\geqslant 0,$$

其中,p_i 表示第 i 个特征的相关度权值,且 $p_i\geqslant 0$.约束条件 $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1$ 是为了避免目标问题出现等于零向量的平凡解.

模型中用 Frobenius 范数约束误差项,主要原因如下.1) 特征选择操作舍弃非相关特征、噪声和冗余特征,保留大部分相关特征,因此拟合误差通常来自于被选择的干净特征.2) 特征选择操作增强模型的鲁棒性,这时使用 Frobenius 范数约束作为损失函数,性能优于 l_1 范数和 $l_{2,1}$ 范数.3) Frobenius 范数具有强凸性和光滑性,在模型优化过程中更易于求解.

命题1[7] 当式(1)满足 EGE 条件1)、2)和4)时,对于所有 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\to 0,\forall k\neq i,k\neq j$,有

1) $|Z_{ii}^*-Z_{jj}^*|\to 0,|Z_{ij}^*-Z_{ji}^*|\to 0$;

2) $|Z_{ik}^*-Z_{jk}^*|\to 0,|Z_{ki}^*-Z_{jk}^*|\to 0$.

由命题1可知,当模型满足 EGE 条件时,若

$$\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\to 0,$$

则

$$\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2\to 0,$$

即在原始数据空间中距离相近的数据点在表示空间中的系数向量仍然相近,称这样的数据点互为强相关.具有光滑性质的迹范数[21]可使每个样本都由其强相关的样本表示,增强矩阵的块对角分布,提升关联矩阵的稀疏性和组效应.通过简单的推导,可得如下正则项:

$$\Omega(\boldsymbol{Z})=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2=\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}}).$$

其中:$\mathrm{tr}(\cdot)$ 为矩阵的迹;$\boldsymbol{W}=(w_{ij})$ 为相似图,$w_{ij}=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2$ 为2个数据点之间的距离相似度,采用0-1权重[22]构造 k 近邻图 $\boldsymbol{W}$;$\boldsymbol{L}$ 为拉普拉斯矩阵,

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W},$$

$\boldsymbol{D}$ 为对角度矩阵,$D_{ii}=\sum_{j=1}^{n}w_{ij}$.至此可得联合特征选择和光滑表示算法的惩罚函数:

$$\min_{\boldsymbol{Z},\boldsymbol{p}} \frac{1}{2}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XZ})\|_{\mathrm{F}}^2+\lambda\,\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}}),\quad (2)$$

$$\text{s.t.}\quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1,p_i\geqslant 0,$$

其中 $\lambda>0$ 为正则项参数.值得注意的是,式(2)中并未限制 $Z_{ii}=0$,因为 $\boldsymbol{X}$ 是由子空间内的样本进行自表示,因此 $Z_{ii}\neq 0$ 有意义;其次,$\lambda>0$ 的设置避免出现如单位矩阵的平凡解[23].

2.2 模型优化方案

目标函数式(2)中系数矩阵 $\boldsymbol{Z}$ 和特征权值 $\boldsymbol{p}$ 相互耦合,一般使用迭代方法求最优解.根据增广拉格朗日方法,引入辅助变量 $\boldsymbol{J}$,令 $\boldsymbol{J}=\boldsymbol{Z}$,则式(2)可转化为

$$L_\rho(\boldsymbol{J},\boldsymbol{Z},\boldsymbol{\Theta})=\frac{1}{2}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XJ})\|_{\mathrm{F}}^2+\lambda\,\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})+\frac{\rho}{2}\left\|\boldsymbol{J}-\boldsymbol{Z}+\frac{\boldsymbol{\Theta}}{\rho}\right\|_{\mathrm{F}}^2,\quad (3)$$

$$\text{s.t.}\quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1,p_i\geqslant 0,$$

其中 $\boldsymbol{\Theta}$ 为拉格朗日算子.引入交替方向乘子法(Alternating Direction Method of Multipliers,ADMM)求解式(3),迭代求取每个目标变量的子问题闭式解.

为了求解 $\boldsymbol{J}$,固定 $\boldsymbol{Z}$、$\boldsymbol{\Theta}$ 和 $\boldsymbol{p}$,可将式(3)改写成

$$\min_{\boldsymbol{J}}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XJ})\|_{\mathrm{F}}^2+\rho\left\|\boldsymbol{J}-\boldsymbol{Z}+\frac{\boldsymbol{\Theta}}{\rho}\right\|_{\mathrm{F}}^2,\quad (4)$$

式(4)是关于 $\boldsymbol{J}$ 的凸优化问题,对其微分置零,即可获得闭式解:

$$\boldsymbol{J}=(\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}^2(\boldsymbol{p})\boldsymbol{X}+\rho\boldsymbol{I})^{-1}\cdot(\boldsymbol{X}^{\mathrm{T}}\mathrm{diag}^2(\boldsymbol{p})\boldsymbol{X}+\rho\boldsymbol{Z}-\boldsymbol{\Theta}),\quad (5)$$

为了求解 $\boldsymbol{Z}$,固定 $\boldsymbol{J}$、$\boldsymbol{\Theta}$ 和 $\boldsymbol{p}$,式(3)可改写成

$$\min_{\boldsymbol{Z}}\lambda\,\mathrm{tr}(\boldsymbol{ZLZ}^{\mathrm{T}})+\frac{\rho}{2}\left\|\boldsymbol{J}-\boldsymbol{Z}+\frac{\boldsymbol{\Theta}}{\rho}\right\|_{\mathrm{F}}^2.\quad (6)$$

类似式(4),式(6)的解为

$$\boldsymbol{Z}=(\rho\boldsymbol{J}+\boldsymbol{\Theta})(2\lambda\boldsymbol{L}+\rho\boldsymbol{I})^{-1}.\quad (7)$$

对于 $\boldsymbol{p}$ 的求解,可固定 $\boldsymbol{J}$、$\boldsymbol{Z}$ 和 $\boldsymbol{\Theta}$,将式(3)改写为

$$\min_{\boldsymbol{p}}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XJ})\|_{\mathrm{F}}^2,$$

$$\text{s.t.}\quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1,p_i\geqslant 0,$$

设

$$q_i^2=\sum_{j=1}^{n}(\boldsymbol{X}-\boldsymbol{XJ})_{ij}^2,$$

该子问题可重写为

$$\min_{\boldsymbol{p}}\|\mathrm{diag}(\boldsymbol{p})(\boldsymbol{X}-\boldsymbol{XJ})\|_{\mathrm{F}}^2=\min\sum_{i}^{d}q_i^2p_i^2=\min\|\boldsymbol{qp}\|_2^2,$$

$$\text{s.t.}\quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1,p_i\geqslant 0.$$

这是一个典型的二次规划问题,常见的二次规划解法有内点法、有效集法和增广拉格朗日法,这里使用增广拉格朗日方法求解该二次规划问题:

$$L(\boldsymbol{p},\eta)=\frac{1}{2}\|\boldsymbol{qp}\|_2^2-\eta(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}-1),$$

其中 η 为拉格朗日算子.在拉格朗日方程中对 $\boldsymbol{p}$ 偏微分置零,可得特征权值的初始值为 $\boldsymbol{p}=(\eta/\boldsymbol{q})_+$,其中,$(a)_+=\max(a,0)$.根据约束条件 $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{1}=1$,可得

$$\sum_{i=1}^{d}\left(\frac{\eta}{q_i}\right)=1,$$

即

$$\eta=\left(\sum_{i=1}^{d}q_i^{-1}\right)^{-1},$$

则 $\boldsymbol{p}$ 的解为

$$\boldsymbol{p}=\left(\frac{1}{\boldsymbol{q}}\left(\sum_{i=1}^{d}q_i^{-1}\right)^{-1}\right)_+.$$

当设定特征选择参数 M 后,对 p_i 降序排序

$$p_1\geqslant p_2\geqslant\cdots\geqslant p_M\geqslant p_{M+1}\geqslant\cdots\geqslant p_d\geqslant 0,$$

当 $i\leqslant M$ 时,

$$p_i=\left(\frac{1}{q_i}\left(\sum_{i=1}^{d}q_i^{-1}\right)^{-1}\right)_+;$$

当 $i>M$ 时,$p_i=0$,定义一个算子 P_M 表示这个关系:

$$\boldsymbol{p}=P_M(\boldsymbol{p}). \tag{8}$$

对于 $\boldsymbol{\Theta}$ 和 ρ 的求解,根据 ADMM 的迭代法则,$\boldsymbol{\Theta}$ 和 ρ 的迭代更新形式可以表示为

$$\boldsymbol{\Theta}=\boldsymbol{\Theta}+\rho(\boldsymbol{J}-\boldsymbol{Z}),\ \rho=\min\left(\rho\kappa,\bar{\rho}\right), \tag{9}$$

其中,$\kappa>1$ 控制收敛速度,$\bar{\rho}$ 用于防止 ρ 变得过大,实验设置为 10^8.综上所述,FSSR 的优化过程如算法 1 描述.

算法 1　联合特征选择和光滑表示算法

输入　数据集 $\boldsymbol{X}$,最大迭代数 $t_{\max}$,参数 λ、ρ、κ

输出　系数矩阵 $\boldsymbol{Z}$

step 1　初始化 $\boldsymbol{Z}$,$\boldsymbol{J}$,$\boldsymbol{p}$,$\boldsymbol{\Theta}$ 为零矩阵或零向量;

step 2　设置迭代次数 $t=1$;

step 3　固定 $\boldsymbol{Z}$、$\boldsymbol{\Theta}$ 和 $\boldsymbol{p}$,根据式(5)更新 $\boldsymbol{J}$;

step 4　固定 $\boldsymbol{J}$、$\boldsymbol{\Theta}$ 和 $\boldsymbol{p}$,根据式(7)更新 $\boldsymbol{Z}$;

step 5　固定 $\boldsymbol{J}$、$\boldsymbol{Z}$ 和 $\boldsymbol{\Theta}$,根据式(8)更新 $\boldsymbol{p}$;

step 6　根据式(9)更新 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\rho}$;

step 7　若 $t\geqslant t_{\max}$ 或收敛,输出结果;否则,$t=t+1$,转至 step 3.

2.3　基于 FSSR 的子空间聚类算法

由算法 1 获得最优的表示系数矩阵 $\boldsymbol{Z}^*$ 后,利用

$$\boldsymbol{S}=\frac{1}{2}(|\boldsymbol{Z}^*|+|\boldsymbol{Z}^{*\mathrm{T}}|), \tag{10}$$

求解关联矩阵 $\boldsymbol{S}$,然后利用规范化分割(Normalized Cut, NCut)方法[10]实现聚类操作.该方法不仅考虑同簇样本的关联性,也兼顾不同簇样本之间的差异.算法 2 给出基于 FSSR 的子空间聚类方法完整描述.

算法 2　基于 FSSR 的子空间聚类算法

输入　数据集 $\boldsymbol{X}$,子空间数 K

输出　样本类标签

step 1　构建 KNN 图 $\boldsymbol{W}$,计算相应的拉普拉斯矩阵 $\boldsymbol{L}$;

step 2　通过算法 1 求出最优系数矩阵 $\boldsymbol{Z}^*$;

step 3　利用式(10)计算关联矩阵 $\boldsymbol{S}$;

step 4　使用谱分割算法获得 K 个子空间.

3　实验及结果分析

3.1　实验设置

将 FSSR 和 5 种聚类方法进行对比,对比算法包括:2 种基于特征选择的方法(CRFS[13]、FSLPS[14]),3 种基于数据表示的方法(SRC[7]、GSLR[9],LRLRR[8]).CRFS、FSLPS 首先根据各自搜索策略选择特征子集,然后使用 k-means 完成最终聚类.SRC、GSLR 和 LRLRR 根据系数矩阵构造关联矩阵,最后使用 NCut 实现数据分簇.

聚类算法通常需要确定使性能指标最高的模型参数,实验中所有对比算法都经过网格搜索[24]选择最优参数取值,表 2 给出预设参数的取值范围.

表 2　算法预设参数的取值范围

Table 2　Range for tunable parameters in competing algorithms

算法	参数设置
CRFS	$k=3,5,\cdots,19,21$; $\lambda=10^{-2},10^{-1},\cdots,10^{3}$; $p=0,0.4,0.6,0.8$
FSLPS	$k=3,5,\cdots,21$; $\gamma=2^{-1},2^{2},2^{4},2^{6},2^{8}$
SRC	$k=3,5,\cdots,21$; $\alpha=10^{-2},10^{-1},\cdots,10^{3}$; $\gamma=10^{-2},10^{-1},\cdots,10^{3}$
GSLR	$k=3,5,\cdots,21$; $p=0.1,0.2,0.4,0.6,0.8,1.0$
LRLRR	$k=3,5,\cdots,21$; $\lambda=10^{-2},10^{-1},\cdots,10^{3}$; $\gamma=10^{-2},10^{-1},\cdots,10^{3}$; $\beta=10^{-2},10^{-1},\cdots,10^{3}$
FSSR	$k=3,5,\cdots,21$; $\lambda=10^{-3},10^{-2},\cdots,10^{3}$

实验使用准确率(Accuracy, ACC)和归一化互信息(Normalized Mutual Information, NMI)[25]作为算法性能的衡量指标.为了保证实验对比的公平性,所有对比算法都进行 10 次独立实验,对比各算法运行结果的平均值.

实验运行条件包括 Intel Core i5 CPU,双核主频 2.60 GHz,内存 4 GB,Windows10 操作系统,

Matlab R2014a 软件.

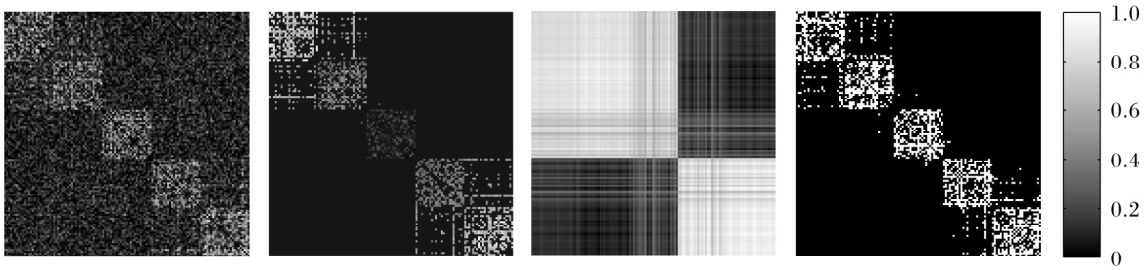
3.2 人造数据实验结果

使用人造合成数据检验 FSSR 的系数矩阵对角效果及其抗噪性能.原始数据由5个间隔0.4的$N(0,1)$高斯分布子空间组成,每个子空间特征维数为250,有25个数据点.选取3种表示型聚类算法 SRC、DRAM、GSLR 与 FSSR 进行对比.图1显示对0%、30%以及60%的特征数据添加$N(0,1)$高斯噪声干扰时,各算法的系数矩阵构造效果.由图1(a)可知,在干净数据上4种算法的系数矩阵都呈现清晰的5类块对角结构,(b)、(c)表明,FSSR 在一定的噪声影响下仍可获得清晰的5类块对角化系数矩阵,其余对比算法的系数矩阵无法保持正确的块对角结构.SRC 因类间存在非零系数而无法完全块对角化,在之后混合噪声情况下更明显.DRAM 系数矩阵的块对角结构较模糊,且不同的类内系数具有不同的均值,不符合聚类分析中数据独立同分布的先验条件.GSLR 在干净数据上获得明显的块对角系数矩阵,但其类内系数固定不变,说明 GSLR 的局部性较差,不适应多模态结构数据.相比之下,FSSR 构造的系数矩阵类间全为零值,类内分布均匀,拥有良好的结构适应性和鲁棒性.

(a1) SRC (a2) DRAM (a3) GSLR (a4) FSSR

(a) 0%高斯噪声

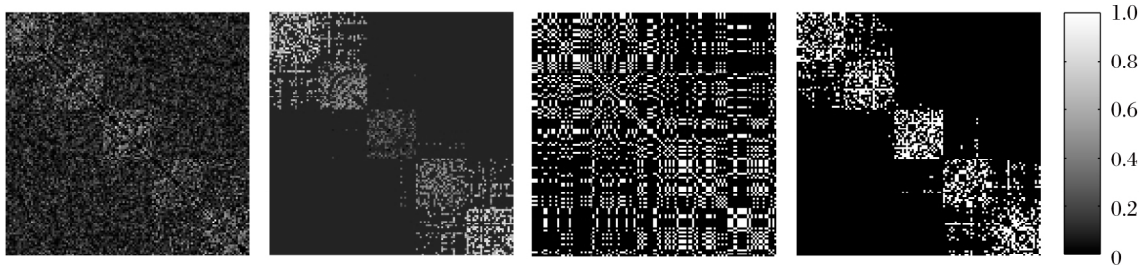
(a) 0% Gaussian noise



(b1) SRC (b2) DRAM (b3) GSLR (b4) FSSR

(b) 30%高斯噪声

(b) 30% Gaussian noise



(c1) SRC (c2) DRAM (c3) GSLR (c4) FSSR

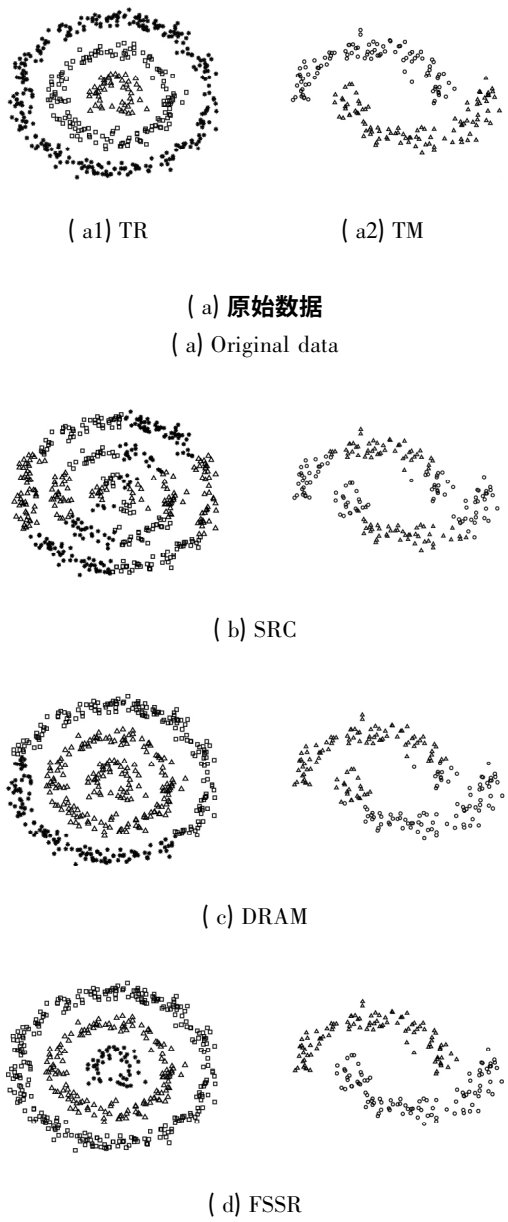
(c) 60%高斯噪声

(c) 60% Gaussian noise

图1 4种算法在人造数据上的系数矩阵

Fig.1 Coefficient matrices of 4 algorithms on synthetic data

使用人造合成的三环(Triple Ring, TR) 数据和双月(Two Moon, TM) 数据模拟验证 FSSR 的聚类效果,如图 2 所示.(a1) 为原始 TR 数据,内侧两环分别具有 150 个数据点,外环具有 200 个数据点.(a2) 为原始 TM 数据,包含 2 类,每类 100 个数据点,坐标交叉放置.(b) ~(d) 为 SRC、DRAM、FSSR 的聚类结果.



(a1) TR　　(a2) TM

(a) 原始数据

(a) Original data

(b) SRC

(c) DRAM

(d) FSSR

图 2　3 种算法在 2 类人造数据上的聚类效果对比

Fig.2　Clustering results comparison of 3 algorithms on 2 synthetic data

由图 2 可见,SRC 和 DRAM 都存在不同程度的聚类错误,而 FSSR 的聚类结果与原始子空间分布完全一致,这证明其优秀的聚类能力.

3.3　标准数据库实验结果

为了验证 FSSR 的实际性能,将其应用在 5 种类型的标准数据库: 3 个人脸数据库(AR、ORL、JAFFE) ,2 个手写字符库(MNIST、USPS) ,1 个医学数据库(LUNG) ,1 个语音数据库(ISOLET) 和 1 个图像数据库(COIL) . 为了体现数据的多样性,选取具有不同特征维度和样本数的数据集,各样本细节信息如表 3 所示.

表 3　标准数据集信息

Table 3　Details of standard datasets

数据集	样本数	特征数	类别数	待选集合
AR	700	792	100	[80,159,238,317,396,476]
ORL	400	1024	40	[103,205,308,410,512,615]
USPS	3000	256	10	[26,52,77,103,128,154]
MNIST	2000	784	10	[79,157,236,314,392,471]
LUNG	203	3312	5	[332,663,994,1325,1656,1988]
JAFFE	213	676	10	[68,136,203,271,338,406]
ISOLET	1559	617	26	[62,124,186,247,309,371]
COIL	1440	1024	20	[103,205,308,410,512,615]

表 4 ~ 表 6 分别为 6 种算法聚类的平均 ACC、NMI 和运行时间,并对每种数据集上排名前三的聚类 ACC 和 NMI 值进行标记,上标 1 为第一名、上标 2 为第二名、上标 3 为第三名.

由表 4 和表 5 可得如下结论:

1) 由于各种数据库数据类型和维度结构具有多样性,每种算法在各个数据库上的聚类效果也呈多样化.ISOLET 语音数据库识别难度较大,各算法聚类效果普遍较低.AR 数据库选用无遮挡的数据集,较易实现较高性能的聚类.

2) 单纯使用特征选择对聚类效果的改善并不明显,而基于自表示模型的算法(SRC,GSLR,LRLRR) 能获得更优的聚类性能.

3) 相比其它算法,FSSR 综合性能具有显著优势,在 ACC 和 NMI 指标上都出现 5 个最高名次,其中在 JAFFE 与 AR 这 2 种数据库上,分别赢得 99.17% 和 93.28% 的 NMI 值,在所有数据库上 FSSR 的 2 种性能指标都在前三名.

最终,FSSR 在所有算法中取得最高的平均 ACC 和 NMI 值,分别为 76.72% 和 80.06%.

表4 6种算法在8个数据集上的聚类ACC对比

Table 4 Clustering ACC comparison of 6 algorithms on 8 datasets

%

算法	AR	ORL	USPS	MNIST	LUNG	JAFFE	ISOLET	COIL
CRFS	34.83	54.29	85.89[1]	57.46[3]	90.56[1]	76.63	57.39[3]	66.16[3]
FSLPS	71.43[3]	47.75	57.90	31.90	80.30	75.12	47.72	60.90
SRC	78.37[2]	73.65[2]	73.17	62.35[2]	89.16[2]	99.03[2]	53.89	69.63[2]
GSLR	43.67	21.25	78.63[2]	51.71	79.93	70.43	29.31	49.58
LRLRR	43.43	63.60[3]	13.60	51.90	71.92	98.13[3]	58.86[1]	62.24
FSSR	82.86[1]	74.00[1]	74.17[3]	65.30[1]	86.70[3]	99.53[1]	58.43[2]	72.78[1]

表5 6种算法在8个数据集上的NMI对比

Table 5 NMI comparison of 6 algorithms on 8 datasets

%

算法	AR	ORL	USPS	MNIST	LUNG	JAFFE	ISOLET	COIL
CRFS	69.93	75.94	84.93[1]	52.82	73.98[1]	89.01	75.32[1]	71.20
FSLPS	91.11[2]	71.56	60.36	39.30	60.54	84.93	70.60[3]	75.98[3]
SRC	90.56[3]	85.11[2]	73.50[2]	61.81[2]	73.57[2]	98.17[2]	69.67	77.93[2]
GSLR	74.59	75.43	69.81	27.68	58.05	78.29	39.60	56.36
LRLRR	74.15	79.14[3]	7.17	53.94[3]	60.20	97.23[3]	70.24	73.04
FSSR	93.28[1]	86.60[1]	72.66[3]	69.53[1]	68.49[3]	99.17[1]	72.38[2]	78.37[1]

表6 6种算法在8个数据集上的运行时间对比

Table 6 Running time comparison of 6 algorithms on 8 datasets

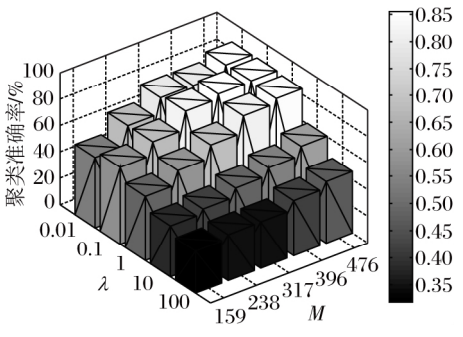
s

算法	AR	ORL	USPS	MNIST	LUNG	JAFFE	ISOLET	COIL
CRFS	19.69	24.58	162.99	98.67	93.39	3.63	40.28	10.17
FSLPS	11.68	2.32	563.28	128.77	5.84	0.40	155.28	489.83
SRC	4.08	0.97	612.80	109.75	0.22	0.14	51.64	37.74
GSLR	2310	445.60	5120	1090	1180	259.7	5310	1650
LRLRR	485.20	143.73	24400	9560	103.23	36.55	4380	3580
FSSR	4.18	2.38	128.69	44.81	1.03	0.33	51.12	48.59

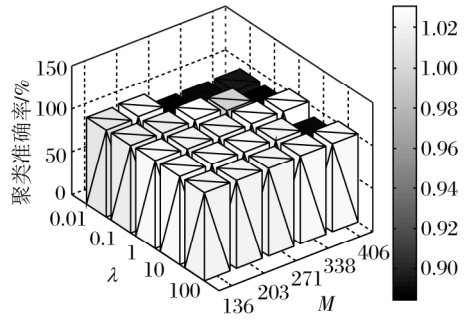
从表6可看出,GSLR和LRLRR在多个数据库上都需千秒以上的运行时间,准确率不高,实际应用价值较低.FSLPS和SRC的运算时间大致属于同一层次,SRC因为不需进行迭代运算而在运行效率上领先于FSLPS.由于对特征选择矩阵的稀疏约束,CRFS具有较高的特征搜索效率,因此平均运行效率略低于本文算法,位居第二.FSSR融合相关特征优选,减少数据处理量,因此获得平均最高的运行效率.

3.4 参数敏感度分析

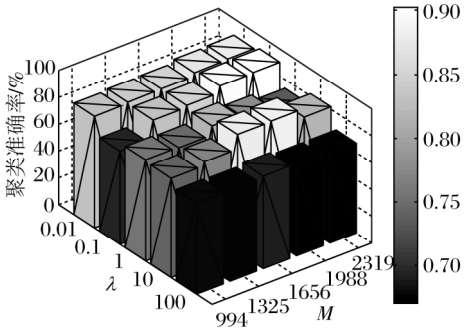
为了确定最优的算法模型,通常对预设参数进行搜索优选,因此,参数的设置个数及其取值敏感度对算法的实际应用效果具有不可忽视的作用.由表3可知,邻域数k是所有无监督聚类算法共有的预设参数,此外,FSSR还有2个重要的预设参数,即正则项参数λ和选择特征数M.图3为在AR、JAFFE、LUNG和COIL等4种数据库上,FSSR的ACC值关于正则项参数和特征选择数的变化关系.



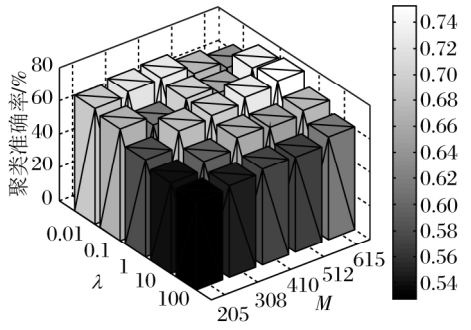
(a) AR



(b) JAFFE



(c) LUNG



(d) ORL

图 3　FSSR 在参数变化下的聚类 ACC

Fig.3　Clustering ACC of FSSR with parameters variation

由图 3 可看出，随着 λ 和 M 的取值变化，FSSR 性能在 JAFFE、LUNG 和 ORL 数据集中都较稳定. 虽然在 AR 数据集上的 ACC 值波动较大，但存在明显变化趋势，这为 FSSR 的实际应用提供直观的可选择性.

4　结 束 语

本文提出联合特征选择和光滑表示的子空间聚类算法(FSSR). 在数据自表示的基础上加入特征选择操作，将无监督特征选择与聚类操作融入同一框架，有效降低高维特征数据聚类的计算复杂度. 通过数据表示框架进行特征权值计算，衡量相关特征对数据表示的贡献程度. 使用迹范数对系数矩阵进行组效应约束，提升系数矩阵的块对角化分布状态. 此外，利用交替方向乘子法迭代更新目标函数各变量的序列值，获得各子问题的闭式求解方案. 通过人造合成数据与标准数据库的对比实验，验证 FSSR 在系数矩阵块对角化、聚类能力和运算效率等方面优于现存同类方法.

通过特征选择向量可以有效删除无效特征，减少噪声点的影响，因此目标函数式(2)仅采用 Frobenius 范数约束误差保真项，可获得良好的聚类性能. 然而，现实数据聚类问题除包含奇异特征点外，还受到奇异样本的困扰. 虽然 Frobenius 范数对高斯噪声具有抑制作用，但无法胜任更复杂的噪声分布类型，后续工作将开展针对样本奇异点的范数形式探索或自适应权值尝试. 此外，通过实验发现，FSSR 对模型参数的变化存在一定程度的敏感性，因此参数自学习或在线更新问题有待更深入研究.

参　考　文　献

[1] AHN I, KIM C. Face and Hair Region Labeling Using Semi-supervised Spectral Clustering-Based Multiple Segmentations. IEEE Transactions on Multimedia, 2016, 18(7): 1414-1421.

[2] LUO J J, JIAO L C, LOZANO J A. A Sparse Spectral Clustering Framework via Multiobjective Evolutionary Algorithm. IEEE Transactions on Evolutionary Computation, 2016, 20(3): 418-433.

[3] VON LUXBURG U. A Tutorial on Spectral Clustering. Statistics and Computing, 2007, 17(4): 395-416.

[4] PENG X, ZHANG L, YI Z. Scalable Sparse Subspace Clustering // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2013: 430-437.

[5] ELHAMIFAR E, VIDAL R. Sparse Subspace Clustering // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2009: 2790-2797.

[6] LIU G C, LIN Z C, YAN S C, *et al*. Robust Recovery of Subspace Structures by Low-Rank Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[7] HU H, LIN Z C, FENG J J, *et al*. Smooth Representation Clustering // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2014: 3834-3841.

[8] YIN M, GAO J B, LIN Z C. Laplacian Regularized Low-Rank Representation and Its Applications. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2016, 38(3): 504-517.

[9] NIE F P, HUANG H. Subspace Clustering via New Low-Rank Model with Discrete Group Structure Constraint // Proc of the 25th International Joint Conference on Artificial Intelligence. Washington,

USA: IEEE, 2016: 1874-1880.

[10] GUO X J. Robust Subspace Segmentation by Simultaneously Learning Data Representations and Their Affinity Matrix // Proc of the 24th International Joint Conference on Artificial Intelligence. Washington, USA: IEEE, 2015: 3547-3553.

[11] CAI D, ZHANG C Y, HE X F. Unsupervised Feature Selection for Multi-cluster Data // Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2010: 333-342.

[12] ZHU P F, ZHU W C, HU Q H, *et al*. Subspace Clustering Guided Unsupervised Feature Selection. Pattern Recognition, 2017, 66: 364-374.

[13] WANG W Z, ZHANG H Z, ZHU P F, *et al*. Non-convex Regularized Self-representation for Unsupervised Feature Selection // Proc of the 5th International Conference on Intelligence Science and Big Data Engineering. New York, USA: ACM, 2015: 55-65.

[14] YAN H, YANG J. Locality Preserving Score for Joint Feature Weights Learning. Neural Networks, 2015, 69: 126-134.

[15] PENG C, KANG Z, YANG M, *et al*. Feature Selection Embedded Subspace Clustering. IEEE Signal Processing Letters, 2016, 23(7): 1018-1022.

[16] 王卫卫,李小平,冯象初,等.稀疏子空间聚类综述.自动化学报, 2015, 41(8): 1373-1384.
(WANG W W, LI X P, FENG X C, *et al*. A Survey on Sparse Subspace Clustering. Acta Automatica Sinica, 2015, 41(8): 1373-1384.)

[17] YONG H W, MENG D Y, ZUO W M, *et al*. Robust Online Matrix Factorization for Dynamic Background Subtraction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. DOI: 10.1109/TPAMI.2017.2732350.

[18] 严 菲,王晓栋.基于局部判别约束的半监督特征选择方法.模式识别与人工智能, 2017, 30(1): 89-95.
(YAN F, WANG X D. A Semi-supervised Feature Selection Method Based on Local Discriminant Constraint. Pattern Recognition and Artificial Intelligence, 2017, 30(1): 89-95.)

[19] CHEN Y, CAO X Y, ZHAO Q, *et al*. Denoising Hyperspectral Image with Non-i.i.d. Noise Structure. IEEE Transactions on Cybernetics, 2017. DOI: 10.1109/TCYE.2017.2677944.

[20] LU C Y, MIN H, ZHAO Z Q, *et al*. Robust and Efficient Subspace Segmentation via Least Squares Regression // Proc of the 12th European Conference on Computer Vision. Berlin, Germany: Springer, 2012: 347-360.

[21] LU C Y, FENG J S, LIN Z C, *et al*. Correlation Adaptive Subspace Segmentation by Trace Lasso // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2013: 1345-1352.

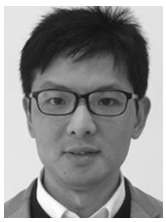
[22] XU Y, ZHONG A N, YANG J, *et al*. LPP Solution Schemes for Use with Face Recognition. Pattern Recognition, 2010, 43: 4165-4176.

[23] PENG C, KANG Z, CHENG Q. Subspace Clustering via Variance Regularized Ridge Regression // Proc of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 682-691.

[24] HUANG C L, WANG C J. A GA-Based Feature Selection and Parameters Optimization for Support Vector Machines. Expert Systems with Applications, 2006, 31(2): 231-240.

[25] ZHAO Z, HE X F, CAI D, *et al*. Graph Regularized Feature Selection with Data Reconstruction. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 689-700.

作者简介



郑建炜,博士,副教授,主要研究方向为机器学习、模式识别. E-mail: zjw@zjut.edu.cn.
(**ZHENG Jianwei**, Ph. D., associate professor. His research interests include machine learning and pattern recognition.)

路 程,硕士研究生,主要研究方向为机器学习、数据挖掘. E-mail: lucheng94@126.com.
(**LU Cheng**, master student. His research interests include machine learning and data mining.)

秦梦洁,硕士研究生,主要研究方向为机器学习、模式识别. E-mail: 1459847947@qq.com.
(**QIN Mengjie**, master student. Her research interests include machine learning and pattern recognition.)

陈婉君(通讯作者),硕士,讲师,主要研究方向为机器学习、模式识别. E-mail: wanjun@zjut.edu.cn.
(**CHEN Wanjun** (Corresponding author), master, lecturer. Her research interests include machine learning and pattern recognition.)